Atomic Energy Control Board
P.O. Box 1046
Ottawa, Canada
K1P 5S9

Commission de contrôle de l'énergie atomique
C.P. 1046
Ottawa, Canada
K1P 5S9

INFO-0264

IMPROVED MORTALITY SEARCHES FOR ONTARIO MINERS USING SOCIAL INSURANCE INDEX IDENTIFIERS

by

M.E. Fair, H.B. Newcombe and
P. Lalonde
Statistics Canada

A research report prepared for the
Atomic Energy Control Board
Ottawa, Canada

February 1988

Canada

Research report

88-06132/

IMPROVED MORTALITY SEARCHES FOR ONTARIO MINERS USING SOCIAL INSURANCE INDEX IDENTIFIERS

A report prepared by Martha E. Fair, Howard B. Newcombe, and Pierre Lalonde, Occupational and Environmental Health Research Unit, Vital Statistics Section, Health Division, Statistics Canada, under contract to the Atomic Energy Control Board.

ABSTRACT

The immediate purpose of the present use of the Social Insurance Number (SIN) index file is to facilitate the death searches pertaining to Ontario miners, including uranium miners. The SIN records contain accurate versions of the personal identifiers such as names and birth dates, whereas these are often incompletely or incorrectly recorded on the available work records such as those of the Workers' Compensation Board (WCB).

The results show that use of the SIN identifiers considerably increases the accuracy of the death searches. Both the false positive and the false negative outcomes from these searches are reduced in number. This is true when the SIN identifiers are employed alone, and even more so when they are used in combination with the WCB identifiers. Moreover, the manual resolutions of the remaining difficult or ambiguous matches are greatly facilitated, in part because of the lessened resemblance between competing death links.

The results are applicable to future studies of the mortality experience of occupational cohorts, including the Newfoundland fluorspar miners and persons enrolled in the National Dose Registry. The improvement will be most marked where SIN registrations for recent years are employed, i.e. following introduction of a requirement to provide birth certificate with the SIN application forms.

As a by-product of the study it has been possible to investigate quantitatively, for the first time, the effect on accuracy of the death searches when various components of the full names, birth dates and such are deleted from the search records. These effects are substantial, and they emphasize further the value of having identifiers of high quality available for any mortality follow-up.

RÉSUMÉ

Dans la présente étude, l'utilisation du fichier des numéros d'assurance sociale vise d'abord à faciliter les recherches concernant les mineurs décédés de l'Ontario, y compris les mineurs d'uranium. Le fichier contient des identificateurs personnels exacts, comme le nom et la date de naissance, tandis que les états nominaux existants de la Commission des accidents de travail, par exemple, sont souvent incomplets ou inexacts.

Les résultats montrent que l'utilisation du fichier des numéros d'assurance sociale accroît considérablement l'exactitude des recherches concernant les mineurs décédés. Il y a moins de résultats faussement positifs ou faussement négatifs de ces recherches lorsque ces identificateurs sont utilisés seuls, et moins encore lorsqu'ils sont utilisés avec les identificateurs de la Commission des accidents de travail. En outre, la résolution manuelle des couplages difficiles ou ambigus s'en trouve grandement facilitée, partiellement en raison de la diminution de la ressemblance entre les couplages concurrents.

Les résultats s'appliquent aux études futures de la mortalité de cohortes professionnelles, y compris les mineurs de spath fluor de Terre-Neuve et les personnes inscrites au Fichier dosimétrique national. L'amélioration se fera surtout sentir lorsque l'on se servira des enregistrements de numéros d'assurance sociale des dernières années, c'est-à-dire depuis qu'il faut joindre un certificat de naissance aux demandes de carte de numéro d'assurance sociale.

L'étude a aussi permis d'étudier quantitativement, pour la première fois, les effets de la suppression, des états nominaux des employés, de divers éléments du nom au complet, de la date de naissance et de données semblables, sur l'exactitude des recherches concernant les mineurs décédés. Ces effets considérables font davantage ressortir l'importance de disposer d'identificateurs de qualité élevée pour tout suivi d'étude de mortalité.

DISCLAIMER

The Atomic Energy Control Board is not responsible for the accuracy of the statements made or opinions expressed in this publication and neither the Board nor the author assumes liability with respect to any damage or loss incurred as a result of the use made of the information contained in this publication. The opinions expressed in this paper are those of the authors and do not necessarily represent the views of Statistics Canada.

PREFACE

The present version of the report on the "SIN evaluation study" has been prepared following production of final data from the relevant death linkages pertaining to Ontario miners. The tables and text of this report have been updated and extended since the draft of March 1986.

The procedures described here are those actually followed, and are not always the same as those discussed tentatively in an earlier planning document (the "Interim Report") which was written prior to initiation of much of the work. Readers wishing to see the details of the earlier tentative plan are referred to:

> "Ontario Miners 'SIN' Evalutation Study -- the 1983/84 Component" Occupational and Environmental Health Research Unit, Vital Statistics and Disease Registries Section, Health Division, Statistics Canada, R.H. Coats Bldg., 18th Floor, Ottawa, Ontario, K1A 0T6, March 1984.

TABLE OF CONTENTS

	PAGE
ABSTRACT	i
PREFACE	iii
A. Introduction	1
B. Available files	1
C. Plan of the study	2
D. Procedures for the study	3
E. Results -- Accuracies of the death links	4
F. Results -- Missing identifiers and the accuracy of the death searches	6
G. Discussion of the value of using SIN index identifiers	6

TEXT TABLES

1. Number of miners' WCB records and number of death records used for the "SIN evaluation" study	8
2. Identifiers on the WCB, SIN and MDB records	9
3. Number of "best" death matches by type of search record -- dead only	10
3a. Number of correct links found, by type of search record	11
4. Matched pairs of records by total weight -- medium breakdown	12
4a. Consequences of choosing various threshold weights for acceptance of a matched pair of records -- medium breakdown	13
5. Matched pairs of records by total weight -- fine breakdown	14
5a. Consequences of choosing various threshold weights for acceptance of a matched pair of records -- fine breakdown	16
6. False positive links and false negative outcomes, using an "optimum" threshold	18
6a. Death matches recognized in the present searches, involving miners previously regarded as "lost to follow-up" -- by weight range, excluding those with weights less than zero	19

TABLE OF CONTENTS - Concluded

PAGE

7. Effects on the error rates when identifiers are deleted from the search records — "hybrid" search records 20

7a. Effects on the error rates when identifiers are deleted from the search records -- "hybrid" search records 22

A. INTRODUCTION

A continuing study of the mortality experience of Ontario miners, including uranium miners, depends for its reliability on the quantity and quality of the personal identifying information contained in the work records that are used to initiate the death searches. However, such information is often incomplete or inaccurate as represented in the Workers' Compensation Board (WCB) records. Better identifiers were believed to exist in the index files of the Social Insurance Number (SIN) system. Use of this more reliable source was therefore considered likely to improve the death searches wherever the Social Insurance Numbers of the miners were known.

The present study was undertaken to test the feasibility of the proposed use of the SIN identifiers (i.e. names, birthdates and such) and to determine how much improvement in the accuracy of the death searches would result.

The procedures were designed to compare the efficiencies of the death searches when initiated respectively by the WCB records alone, the SIN records alone, and by composite or "hybrid" records containing identifiers from both. A file of known correct death links was used in the test. Comparisons of the error rates for the present searches, when initiated by the three types of records, were based on this prior verification of the potential good links.

The anticipated results from the present tests were thought to have numerous applications in virtually any future epidemiological investigations of mortality in groups for which SIN numbers are available. Other examples of such groups include persons enrolled in the National Dose Registry and the Newfoundland fluorspar miners.

In addition, the tests were designed to yield, as a relevant byproduct, quantitative data on the extents to which the absences of particular identifiers or groups of identifiers act to reduce the accuracy of the death searches.

B. AVAILABLE FILES

The study cohort of Ontario miners is represented by about 50,279 records from the Workers' Compensation Board (WCB). Of these, a total of precisely 30,000 contain valid Social Insurance Numbers (SINs) and are therefore numerically linkable with the records of the SIN index (see Table 1). Records from both these sources are potentially linkable with the death records contained in Canada's Mortality Data Base (MDB), using names and such.

The vital status of the 30,000 miners whose SIN numbers are known is also indicated in Table 1. This had been ascertained earlier by a particularly thorough process consisting of: a) a death search, b) an "alive" search of the taxation files, followed by c) a manual check of any apparent conflicts to uncover the reason and the true status.

The computerized death searches are probabilistic in nature. Although the process is based primarily on names and birth dates, these are best complemented wherever possible with such additional identifiers as the place of birth, the place of residence or of work, the names of relatives (e.g. mother's maiden surname), and the year when the person was last known to be alive. The extents to which provision is made for such linking information

on the above three sorts of records (WCB, SIN, and MDB) is shown in Table 2.

Linkage of the WCB records with those of the SIN index is technically simple using the Social Insurance Number. Because the WCB records lack the mother's maiden surname which is on the SIN records, and because the SIN records lack the province or foreign country of birth which is sometimes available from the WCB records, the two together will often contain more identifiers than either alone. Moreover, where the identifiers are differently recorded on the WCB and SIN records, the two together will be more likely to contain the version given on the death record than would either of them alone. Both of these considerations should improve the capacity to establish the correct death links, where a combined or "hybrid" (WCB-SIN) record is employed to initiate the searches of the Mortality Data Base.

C. PLAN OF THE STUDY

Two prior linkage operations greatly facilitated the present "SIN evaluation" study. These are: a) the death searches up to 1977 already initiated by the WCB records, and b) the "alive follow-up" carried out earlier in which the WCB records that have SIN numbers were numerically linked with the income tax records for 1977 and 1978. The former undertaking greatly reduced the numbers of death records that needed to be handled in order to compare the efficiencies of the death searches with and without the aid of the SIN index records. The latter operation had already served to confirm the vital status of the miners where the results from the death searches alone would have been ambiguous or wrong.

From the prior "death" searches, at an intermediate stage in the linkage operation, a comparatively small file of death records was established which included all that were even remotely likely to link with the miners' records either correctly or falsely (see the "prod7-DATB" files in Table 1). All 30,000 of the miners' records shared a surname NYSIIS code with one or more of the records in this truncated death file.

From the prior "alive" searches, the vital status had been clearly established for the 30,000 miners whose WCB records contained a SIN number. Of these, a) 26,736 were confirmed "alive" at the end of 1977, b) 2,254 were confirmed "dead" at that time, and c) 1,010 were designated "lost to follow-up" (see Table 1). "Lost to follow-up" means that their names were absent from both the 1977 and 1978 tax files and their WCB vital status was unknown. (One WCB record with a valid SIN number was excluded because the number was found to have been used by two different people. This was the 30,001st WCB record, and it was simply removed for the purposes of this study.)

The plan for the present study was to repeat the death searches using,

1. the WCB identifiers alone,
2. the SIN index identifiers alone, and
3. both sets of identifiers.

Use of both sets of identifiers (3, above) implies the creation of composite or "hybrid" records. Such records were designed to contain the mother's maiden surname from the SIN record because it is not recorded in the WCB record, and the birth place code from the WCB record because it is absent from

the SIN record (see Table 2). The hybrid records were made to include also any other identifier that happened to be present on either source record. Where a
given identifier (e.g. a name or a birth date) was represented differently on the two source records, then two separate "entry records" were created for the individual miner. These were the same except for the discrepant identifier (or identifiers).

The results from the above three kinds of death searches are best judged in terms of the numbers of linkage errors (false positives and false negatives) arising from each. Some of the 2254 known potential good links were not represented by matched pairs in the present searches. These "missed linkages" have been included, however, in the numerators and denominators on which the final error rates were calculated.

Conversely, some of the 1010 miners' records belonging to the "lost to follow-up" category did form matched pairs with the death records as a result of the present searches. A small number of these pairs do carry high total weights and probably represent valid death links. However, because the vital status of these miners had not been verified earlier, the pairs in question cannot be said definitely to be "good" links or "bad" links. For this reason they are not employed in calculating the error rates. Their numbers, however, probably say something less precise about the abilities of the three types of search records to "capture" the relevant death records.

The plan included the capture of additional information concerning the importance of missing identifiers as a source of error in any linkage operation. Birth dates and given names are known to be particularly important, but in the past there have been no quantitative data on the magnitudes of the errors created when various identifiers are dropped from the records prior to linking them. Such data were not too difficult to obtain as a byproduct.

D. PROCEDURES FOR THE STUDY

For the present study use was made of four files, a death file plus three sets of records pertaining to Ontario miners. The latter were used to initiate the death searches. More specifically, these files consisted of:

1. Death records from the MDB that had survived the "COMPARE" phase of an earlier death search,
2. WCB records with SIN numbers,
3. SIN index records with the same SIN numbers as the WCB records, and
4. Composite WCB-SIN records from the above.

The death records (1, above) were from a file created at an intermediate stage in the previous computerized search of the Mortality Data Base; the file is known as "PROD7-OUTDATB" (or "PROD7-DATB", see Table 1). This file contains all death records for which there was even a remote possibility that they might link with a WCB record. Death records that were excluded from this file are highly unlikely to be capable of forming even false linkages, and so may be regarded as irrelevant for present purposes.

Of the 30,000 WCB records with SIN numbers, only 2,254 related to miners who were actually known to be dead. However, all 30,000 have been included in the test since some of them could conceivably contribute to the numbers of false linkages.

For each WCB record used in the test, the corresponding record with the same SIN number was drawn from the SIN index file. These records were used, both separately (2 and 3, above) and together as composite or "hybrid" WCB-SIN records (4, above).

The procedures for carrying out probabilistic death searches, based on names, birth dates, places of birth, places of residence and such, have now become routine at Statistics Canada. Readers who are interested in the details as relating to the present study are referred to the interim report ("Ontario Miners 'SIN' Evaluation Study -- The 1983-84 Component") mentioned in the Preface. The principle is simple enough. Where an identifier on a matched pair of records agrees or disagrees (or is otherwise similar or dissimilar), one simply asks how common this particular comparison "outcome" is among linked pairs of records as compared with unlinkable pairs brought together at random. This ratio represents the extent to which the overall odds in favour of linkage will be affected, in one direction or the other, by the comparison outcome. It is just a matter of common sense that outcomes which are more "fashionable" among linked pairs (e.g. agreements) will argue for linkage, and that those which are more characteristic of unlinked pairs (e.g. disagreements) will argue against linkage. It is convenient to express these ratios as logarithms so that they can be added to each other, and logs to the base 2 are often used because they can be calculated approximately on one's fingers. But these are details that should not be allowed to obscure the intuitively obvious principle that underlies the probabilistic death searches and linkages.

The only special feature of the present death searches arises where, for a particular identifier, the WCB-version and the SIN-version of that identifier differ. In that situation two composite or "hybrid" records were created and both of these instituted essentially independent searches of the death file. The better of the two possible death matches was then selected. This, of course, did not exclude the possibility of both matches being examined visually during a "manual resolution" of any doubtful pairings.

To avoid confusion later, the reader is advised that the overall odds will be represented in the Tables of this report as a logarithm to the base 2, multiplied by 10 so that decimals do not need to be used. A total weight of zero may thus be taken as representing calculated 50 : 50 odds, a total weight of + 10 represents odds of 2 : 1 in favour of linkage, a total weight of - 10 represents odds of 2 : 1 against, and so on.

E. RESULTS -- ACCURACIES OF THE DEATH LINKS

The study was designed primarily to answer a simple question, i.e. "Does use of the SIN identifiers substantially improve the accuracy of the death links?" The simple answer will be in terms of the potential good links that are missed. The numbers of such false negatives are expected to be reduced where the SIN identifiers are employed in the death searches, either alone or in combination with the WCB identifiers. The more precise but less simple answer, however, must also take into consideration the numbers of false links that get accepted, and the rules by which one decides whether to accept or reject matched pairs when one does not know in advance which are correct and which are false. These rules will have to do with the choice of a "threshold" weight or level of calculated odds, below which the pairs are rejected.

The simple answer to the above question is that use of the SIN identifiers greatly increased the capture of the known potential correct death links (see Table 3). With the WCB identifiers alone, 56 of the 2254 potential good links were missed entirely; with the use of the SIN plus WCB identifiers the number of such losses was reduced to just 7 (i.e. 0.3 per cent as against 2.5 per cent). But this does not tell us about the numbers of false positive links that would have been accepted by the computer, and we already know that for a few of the correct pairings there was a competing falsely matched pair that would have been preferred by the computer on the basis of the total weight.

The more precise but less simple answer to the question requires that we consider the use of some "optimum" threshold weight for acceptance or rejection of the matched pairs. For present purposes the "optimum" may be defined as the level which ensures that the false positives and false negatives come closest to being equal in number. With this "optimum" threshold, the total errors (i.e. false positives plus false negatives) tend to be reasonably close to a minimum; i.e. the total number of errors usually rises if one moves the threshold substantially above or below the "optimum". In setting such a threshold, we have to ignore for the moment the known potential good links that have failed, for reasons other than the total weight, to give rise to matched pairs in the present operation.

To find this "optimum" threshold, the matched pairs of records (i.e. the best death match for each miner's record) are arrayed in descending order of total weight. We have used a breakdown by ranges of 10 units of weight (Tables 4), and a finer breakdown by single units of weight for those weights in the vicinity of zero (Tables 5). From these distributions, a precise "optimum" threshold may be derived for the death searches initiated in each of the three ways (Tables 4a and 5a). Such "optimum" thresholds may vary with the amount of identifying information available. They also tend to fall somewhat above the theoretical 50 : 50 odds point (total weight = zero), presumably because of the large number of falsely matched pairs which are candidates for inclusion and which may sometimes achieve total weights above zero, especially when they are numerous to start with.

Using the "optimum" thresholds so derived, one may determine the numbers of false linkages that are accepted because the total weights are above that threshold, and the numbers of potential good matches that are rejected because the total weights are below the threshold. For each of the three kinds of death search, these two sorts of false outcomes may be expressed as percentages of the total potential correct death linkages as determined prior to this study (Table 6). The final percentages should, of course, include the known potential good links that failed to be represented by matched pairs in the present searches (see lower portion of Table 6).

The results indicate that the SIN records alone are clearly superior to the WCB records alone. With the SIN records, more of those that were potentially linkable succeeded in forming correct pairs with the death records, more of the truly linkable pairs achieved total weights above the "optimum" threshold, and fewer of the truly unlinkable pairs did so (Table 6, top two lines of each section).

The composite or "hybrid" records were better than either the SIN or the WCB records alone. With these, the total error rate was approximately 5 per cent of the potentially linkable pairs, as compared with just under 10 per cent for

the WCB records alone (Table 6, see section "based on all search records"). Again, more of the potentially linkable records succeeded in forming pairs with the death records, more of the truly linkable pairs that were formed were accepted, and fewer of the truly unlinkable pairs were. There is also evidence that the "hybrid" records pertaining to miners with a previous "lost to follow-up" status may have established valid death linkages in the present searches, in larger number than is the case for the corresponding WCB records. Again, the SIN records alone are intermediate between the WCB and the "hybrid" records (see Table 6a).

The advantages of using the personal identifiers contained in the SIN index files are substantial and obvious. For those who may still wonder why even higher levels of accuracy of the death searches are not readily achieved, the answer lies in the adequacy of the identifying information contained in the original records. Some of the identifying information gets wrongly recorded, and some does not get recorded at all. The effect of the missing identifiers will be considered in the next section.

F. RESULTS — MISSING IDENTIFIERS AND THE ACCURACY OF THE DEATH SEARCHES

Finally, it is possible to investigate the extents to which the error rates are influenced by the availability or non-availability of the various identifiers on the search records. This is best done by suppressing certain of the more important identifiers contained in the search records (e.g. the first name in full, a middle initial, the date of birth in full, and the place of birth), singly and in groups. The death searches are then repeated, with these present and with certain of them removed or not used. The results from a series of such tests are shown in Tables 7 and 7a.

These tests show clearly that use, or non use, of even a seemingly less important identifier such as a middle name or initial (which is often not present anyway) still makes a major difference to the accuracy of the death searches.

Wherever forms are being filled out, the need to record certain of the personal identifiers is constantly open to review by those who wish to minimize the labour and the possible unfavourable responses from the informants. However, where these have to be balanced against a need to use the records for later file searches, the penalty for omission of a particular identifier should be recognized. The present data demonstrate the magnitude of that penalty in terms of reduced accuracy of the subsequent file searches when initiated by the records in question. Clearly, the full date of birth is of special importance in this connection, and the full given name comes next to it. Even the less frequently used identifiers such as the place of birth and the mother's maiden surname seem to become increasingly important where there is a shortage of the more frequently used identifiers.

G. DISCUSSION OF THE VALUE OF USING SIN INDEX IDENTIFIERS

The present test provides a measure of the usefulness of employing personal identifiers from the SIN index to supplement those from the work records, for the purpose of searching the Mortality Data Base. The improvement in the accuracy of the automated linkages is obvious. What is less obvious is the

degree to which the manual resolutions and verifications of difficult links are also facilitated.

Two effects deserve special mention. First, the SIN records have more accurate surnames. This means that failures to bring potentially linkable records together for comparison, due to differences that affect the surname phonetic (NYSIIS) code, are much less frequent where the SIN identifiers are used. In the present study the "hybrid" search generated an additional 49 links, i.e. 2.2 per cent, raising the capture rate from 97.3 to 99.5 per cent of the potential good links. Second, the other identifiers are likewise more accurately and more completely reported on the SIN records. This ensures that the good links out-distance the runner-up false links by a considerably wider weight margin than they would with WCB identifiers alone. Both effects help the computer to arrive at a correct judgment. They help to an even greater degree the process of visual assessment and manual resolution.

The tests to determine the effects on the accuracy of searching, when various personal identifiers are deleted, serve to emphasize the value of using a source of these identifiers (the SIN files) in which they are of high quality and completeness.

It is not just the Ontario miners study that could benefit from use of the identifiers contained in the SIN records. These would be of value in virtually any investigation of the mortality experience of occupational groups for whom the SIN numbers are known. This is true in particular of the follow-up of persons enrolled in the National Dose Registry, and of the Newfoundland fluorspar miners. Moreover, for the future there are likely to be many other occupational groups for whom the death searches could be similarly improved.

Table 1. Number of Miners' WCB Records and Number of Death Records Used for the "SIN Evaluation" Study

Records	Totals
MINERS' WCB RECORDS (MALES)	
Total in original file	50,279
Total with valid SIN numbers	30,000
Vital status of the 30,000 miners with SIN numbers:	
Confirmed alive	26,736
Death links (1964-77) confirmed	2,254
Lost to follow-up	1,010
DEATH RECORDS SEARCHED (MALES)	
Total in MDB (1964-1977) approx.	1,300,000
Total in "prod7-DATB" file (1950-1977)	46,679
Total in "prod7-DATB" file (1964-1977)	35,251

Table 2. Identifiers on the WCB, SIN and MDB Records

Identifier	WCB Rec	SIN Rec	MDB Rec
Social Insurance Number	+	+	n.a.
Names*	+	+	+
Birth date	+	+	+*
Birth place	+**	n.a.	+*
Sex	+	+	+
Province of residence	+	n.a.	+
Mother's maiden surname	n.a.	+	+*
Year last known alive	+	n.a.	+

\+ Present on the file

* Not always complete.

** Canada and the U.S.A. grouped together under the same code -- not very discriminating.

n.a. - not available

Table 3. Number of "Best" Death Matches by Type of Search Record — Dead Only

(The searches all relate to the 2,254 miners known to have died in the period 1964-77.)

Kind of Match Achieved	Number of Search Records		
	WCB Search Record	SIN Search Record	Hybrid Search Record
COUNTS			
Correct match was made; it was the "best"	2,193	2,238	2,243
Correct match made; it was not the "best"	5	3	4
No match was made	56	13	7
Total search records	2,254	2,254	2,254
PERCENTAGES			
Correct match was made; it was the "best"	97.29	99.29	99.51
Correct match made; it was not the "best"	.22	.13	.18
No match was made	2 .48	.58	.31
Total search records	100.00	100.00	100.00

NOTE: The "best" match is by definition the one with the highest calculated odds in favour of a correct death linkage.

Table 3a. Number of Correct Links Found, by Type of Search Record (Based on the same data as Table 3.)

Correct Links Found by					
WCB Search Record	SIN Search Record	Hybrid Search Record		Number of Correct Links Found*	Percentage of Correct Links Found*
Yes	Yes	Yes		2,192	97.25
No	Yes	Yes		49	2.17
Yes	No	Yes		6	.27
No	No	No		7**	.31
			Total	2,254	100.00
COMBINED					
Yes	--	--		2,198	97.52
--	Yes	--		2,241	99.42
--	--	Yes		2,247	99.69

* All correct links are included, including the few that carried lower weights than an alternative incorrect death match.

** The 7 potential correct death links that were not detected by any of the three kinds of search records included:

4 due to differences in the surname NYSIIS code,
2 due to late registration of the death,
1 for reasons unknown.

Table 4. Matched Pairs of Records by Total Weight — Medium Breakdown

(Including only the "best" death match for each miner's record.* Based on 28,990 search records, i.e. excluding 1,010 "lost to follow-up" records some of which formed pairs in this last linkage run that have not yet been manually resolved.)

Total weight		WCB search		SIN search		Hybrid search	
odds ratio	wt range	good pairs	bad pairs	good pairs	bad pairs	good pairs	bad pairs
$2^{11}+$	105 and over	1,760	8	2,006	8	2,091	15
2^{10}	95 to 104	60	1	23	2	21	4
2^{9}	85 94	53	3	28	3	22	5
2^{8}	75 84	50	3	28	4	15	4
2^{7}	65 74	40	4	29	2	20	9
2^{6}	55 64	49	8	19	5	11	9
2^{5}	45 to 54	35	-	16	2	17	15
2^{4}	35 44	27	13	15	6	7	16
2^{3}	25 34	19	16	12	13	9	21
2^{2}	15 24	17	22	9	13	7	26
2^{1}	5 14	18	27	10	18	2	26
$2^{0}+$	0 to 4	12	15	3	8	4	26
$2^{0}-$	- 5 - 1	5	18	5	15	5	22
2^{-1}	-15 to - 6	17	61	6	43	2	59
2^{-2}	-25 -16	5	97	5	51	1	75
2^{-3}	-35 -26	6	97	6	68	2	91
3^{-4}	-45 -36	7	153	4	65	1	137
2^{-5}	-55 -46	3	188	4	104	-	156
2^{-6}	-65 to -56	3	216	1	139	3	213
2^{-7}	-75 -66	3	230	3	179	1	223
2^{-8}	-85 -76	2	262	-	204	2	273
2^{-9}	-95 -86	-	309	1	245	-	268
2^{-10}	-105 -96	1	214	-	191	-	211
2^{-11}	-115 and below	1	453	5	618	-	560
no match at all		56	24,318	13	24730	7	24,272
correct but not "best"*		5		3		4	
Total		2,254	26,736	2,254	26,736	2,254	26,736

* The "best" match is by definition the one with the highest calculated odds in favour of a correct death linkage.

Table 4a. Consequences of Choosing Various Threshold Weights for Acceptance of a Matched Pair of Records — Medium Breakdown

(Based on the same data as Table 4.)

Threshold weight for acceptance	WCB search		SIN search		Hybrid search	
	false neg	false pos	false neg	false pos	false neg	false pos
500	2,193	-	2,221	-	2,199	-
400	2,165	-	2,052	-	1,999	-
300	1,989	-	1,517	-	1,383	-
200	1,283	1	738	1	569	2
150	763	1	434	1	292	3
100	398	8	215	9	141	18
90	347	10	195	10	116	21
80	290	15	171	16	100	26
70	253	16	143	18	83	33
60	206	24	115	22	68	43
50	163	27	92	25	55*	54*
40	132	35	79	29	41	66
30	110	47	65	37	32	86
20	95*	67*	56*	52*	25	108
10	74*	90*	46	66	22	132
0	53	120	40	84	17	176
- 9	35	155	32	118	12	221
-19	30	239	28	168	10	284
-29	23	329	20	214	8	366
-39	17	443	17	286	6	471
-49	12	619	10	360	6	611
-59	8	819	10	476	4	801
-69	7	1,053	9	641	3	1,013
-79	2	1,275	6	826	-	1,257
-89	2	1,567	6	1,053	-	1,533
-99	1	1,836	5	1,284	-	1,784
-109	1	2,071	3	1,497	-	2,021
-119	-	2,230	1	1,690	-	2,222
-129	-	2,300	-	1,836	-	2,329
-139	-	2,349	-	1,914	-	2,390
-149	-	2,377	-	1,954	-	2,416
-199	-	2,418	-	2,006	-	2,464

* "Optimum" threshold.

Table 5. Matched Pairs of Records by Total Weight — Fine Breakdown

Total weight	WCB search		SIN search		Hybrid search	
	good pairs	bad pairs	good pairs	bad pairs	good pairs	bad pairs
25+	2,093	56	2,176	45	2,213	98
24	-	3	2	2	-	4
23	1	1	1	3	1	2
22	2	2	1	1	-	1
21	1	1	1	-	1	1
20	1	4	1	1	3	2
19	2	1	-	1	1	2
18	2	4	1	1	-	3
17	4	3	1	1	1	6
16	4	1	-	1	-	2
15	-	2	1	2	-	3
14	5	3	1	-	-	2
13	3	4	1	1	-	1
12	1	1	1	1	1	-
11	-	-	2	1	-	3
10	-	4	2	5	-	2
9	-	2	1	3	-	2
8	1	1	-	3	1	3
7	-	4	-	-	-	6
6	5	4	-	1	-	4
5	3	4	2	3	-	3
4	3	1	-	-	1	4
3	1	4	-	2	1	7
2	7	3	1	-	-	7
1	-	3	2	2	-	5
0	1	4	-	4	2	3
-1	-	2	1	4	1	1
-2	2	-	1	4	1	4
-3	1	5	-	3	1	7
-4	2	7	2	1	1	7
-5	-	4	1	3	1	3
-6	5	3	-	4	-	8
-7	3	4	1	5	-	4
-8	5	3	1	8	-	5
-9	-	7	1	2	-	6
-10	-	5	1	2	-	7

Table 5. Matched Pairs of Records by Total Weight — Fine Breakdown - Concluded

Total weight	WCB search		SIN search		Hybrid search	
	good pairs	bad pairs	good pairs	bad pairs	good pairs	bad pairs
-11	-	8	-	5	1	8
-12	2	7	-	4	-	7
-13	1	5	-	3	-	4
-14	1	10	1	5	1	5
-15(-)	31	2,228	30	1,869	10	2,212
no match at all	56	24,318	13	24,730	7	24,272
correct but not "best"	5		3		4	
Total	2,254	26,736	2,254	26,736	2,254	26,736

Table 5a. Consequences of Choosing Various Threshold Weights for Acceptance of a Matched Pair of Records — Fine Breakdown

(Based on the same data as Table 5.)

Threshold weight for acceptance	WCB search		SIN search		Hybrid search	
	false neg	false pos	false neg	false pos	false neg	false pos
24	100	59	60	47	30	101
23	99	60	59	50	29	104
22	97	62	58	51	29	105
21	96	63	57	51	28	106
20	95	67	56	52	25	108
19	93	71	56	53	24	110
18	91	72	55*	54*	24	113
17	87	76	54*	55*	23	119
16	83	79	54	56	23	121
15	83*	80*	53	58	23	124
14	78	82	52	58	23	126
13	75	85	51	59	23	127
12	74	89	50	60	22	127
11	74	90	48	61	22	130
10	74	90	46	66	22	132
9	74	94	45	69	22	134
8	73	96	45	72	21	137
7	73	97	45	72	21	143
6	68	101	45	73	21	147
5	65	105	43	76	21	150
4	62	109	43	76	20	154
3	61	110	43	78	19	161
2	54	114	42	78	19	168
1	54	117	40	80	19	168
0	53	120	40	84	17	176
-1	53	122	39	88	16	177
-2	51	122	38	92	15	181
-3	50	127	38	95	14	188
-4	48	134	36	96	13	195
-5	48	138	35	99	12	198
-6	43	141	35	103	12	206
-7	40	145	34	108	12	210
-8	35	152	33	116	12	215
-9	35	155	32	118	12	221
-10	35	160	31	120	12	228

Table 5a. Consequences of Choosing Various Threshold Weights for Acceptance of a Matched Pair of Records — Fine Breakdown - Concluded

Threshold weight for acceptance	WCB search		SIN search		Hybrid search	
	false neg	false pos	false neg	false pos	false neg	false pos
-11	35	168	31	125	11	236
-12	33	175	31	129	11	243
-13	32	180	31	132	11	247
-14	31	190	30	137	10	252

* "Optimum" threshold.

Table 6. False Positive Links and False Negative Outcomes, Using an "Optimum" Threshold

Search records (number potentially linkable)	Optimum threshold	False pos		False neg		Total	
		No	(%)	No	(%)	No	(%)
BASED ON SEARCH RECORDS THAT FORMED MATCHED PAIRS (Excludes the known truly linkable search records that did not form matched pairs in this operation.)							
WCB records (2198)	+15	80	(3.6)	83	(3.8)	163	(7.4)
SIN records (2241)	+16	55	(2.5)	54	(2.4)	109	(4.9)
Hybrid rec. (2247)	+50	55	(2.4)	54	(2.4)	109	(4.9)
BASED ON ALL SEARCH RECORDS (Includes the known truly linkable search records that did not form matched pairs in this operation.)							
WCB records (2254)	+15	80	(3.5)	139	(6.2)	219	(9.7)
SIN records (2254)	+16	55	(2.4)	67	(3.0)	122	(5.4)
Hybrid rec. (2254)	+50	55	(2.4)	61	(2.7)	116	(5.1)

Table 6a. Death Matches Recognized in the Present Searches, Involving Miners Previously Regarded as "Lost to Follow-up" — by Weight Range, Excluding those with Weights Less than Zero

Weight range	Matched pairs of records		
	WCB search	SIN search	Hybrid search
100 and over	4	10	11
50 to 99	2	2	5
20 to 49	6	3	3
10 to 19	4	1	4
0 to 9	1	3	2
Total zero and over	17	19	25

Table 7. Effects on the Error Rates when Identifiers are Deleted from the "Hybrid" Search Records[1]

RUN #	Birth Yr	Birth Mo	Birth Dy	Given Names I1	Given Names R1	Given Names I2	Given Names R2	MM	BP	Best Threshold	Counts False Pos	Counts False Neg	Total Error	Total Errors as a % of 705 Good Link
DELETIONS OF BIRTH DATE INFORMATION														
1.+	+	+	+	+	+	+	+	+	.	50+	8	8	16	2.3
2.+	+	-	+	+	+	+	+	+	.	46+	12	12	24	3.4
3.+	-	-	+	+	+	+	+	+	.	50+	23	23	46	6.5
4.-	-	-	+	+	+	+	+	+	.	37+	45	45 + 3*	93	13.2
5.+	+	+	+	+	+	+	+	-	.	37+	14	14	28	4.0
6.+	+	-	+	+	+	+	+	-	.	25+	18	19	37	5.2
7.+	-	-	+	+	+	+	+	-	.	24+	33	34	67	9.5
8.-	-	-	+	+	+	+	+	-	.	4+	76	78 + 9*	163	23.1
DELETIONS OF NAME INFORMATION														
1.+	+	+	+	+	+	+	+	+	.	50+	8	8	16	2.3
9.+	+	+	+	+	-	-	+	+	.	45+	15	14	29	4.1
10.+	+	+	+	-	-	-	+	+	.	45+	20	20	40	5.7
5.+	+	+	+	+	+	+	+	-	.	37+	14	14	28	4.0
11.+	+	+	+	+	-	-	-	-	.	41+	20	21 + 1*	42	6.0
12.+	+	+	+	-	-	-	-	-	.	7+	25	25 + 1*	51	7.2

* See Footnotes for Tables 7 and 7a

[1] Based on just those search records from the correctly matched pairs having a full set of identifiers on both records — i.e. full date of birth, two full given names, and mother's maiden surname. Birthplace was present or absent on both records. Percentages are calculated on 705 potential good links.

Key to Headings:
I - Initials of first (I1) or second (I2) given name
R - Remainder of first (R1) or second (R2) given name
MM - Mother's maiden surname
BP - Birth place

Key to Entries:
Identifier: Deleted (-); Present (+); Present or absent (.)

Table 7. (continued)
Effects on the Error Rates when Identifiers are Deleted from the "Hybrid" Search Records

RUN #	Birth Yr	Birth Mo	Birth Dy	Given Names I1	Given Names R1	Given Names I2	Given Names R2	MM	BP	Best Threshold	Counts False Pos	Counts False Neg	Total Error	Total Errors as a % of 705 Good Link
DELETIONS OF MOTHER'S MAIDEN SURNAME														
1.+	+	+	+	+	+	+	+	+	.	50+	8	8	16	2.3
5.+	+	+	+	+	+	+	+	-	.	37+	14	14	28	4.0
DELETIONS OF BIRTHPLACE														
1.+	+	+	+	+	+	+	+	+	.	50+	8	8	16	2.3
13.+	+	+	+	+	+	+	+	+	-	49+	8	8	16	2.3
MULTIPLE DELETIONS OF IDENTIFIERS														
1.+	+	+	+	+	+	+	+	+	.	50+	8	8	16	2.3
3.+	-	-	+	+	+	+	+	+	.	50+	23	23	46	6.5
15.+	-	-	+	+	-	-	+	.		44+	39	41 + 1*	81	11.5
16.+	-	-	+	+	-	-	-	.		41+	40	41 + 2*	83	11.8
17.+	-	-	+	+	-	-	-	-		1+	70	60 + 5*	135	19.1
20.+	-	-	+	-	-	-	-	.		-16+	105	121 + 8*	234	33.2
18.-	-	-	+	+	-	-	+	.		21+	87	89 + 10*	186	26.4
19.-	-	-	+	+	-	-	-	.		-14+	122	162 + 32*	316	44.8
14.+	+	+	+	+	+	+	-	-		36+	13	13 + 1*	27	3.8

* See Footnotes for Tables 2 and 3

Key to Headings:
I - Initials of first (I1) or second (I2) given name
R - Remainder of first (R1) or second (R2) given name
MM - Mother's maiden surname
BP - Birth place

Key to Entries:
Identifier: Deleted (-); Present (+); Present or absent (.)

Table 7a. Effects on the Error Rates when Identifiers are Deleted from the "Hybrid" Search Records[1]

RUN #	Birth Yr	Birth Mo	Birth Dy	Given Names I1	Given Names R1	Given Names I2	Given Names R2	MM	BP	Best Threshold	Counts False Pos	Counts False Neg	Counts Total Error	Total Errors % of 2243 Good Links
DELETIONS OF BIRTH DATE INFORMATION														
1.	.	.	.	.	.	.	.	.	.	50+	54	55	109	4.9
2.	.	.	-	.	.	.	.	.	.	40+	82	81 + 1*	164	7.3
3.	.	-	-	.	.	.	.	.	.	38+	125	123 + 10*	258	11.5
4.	-	-	-	.	.	.	.	.	.	16+	229	233 + 41*	503	22.4
DELETIONS OF NAME INFORMATION														
1.	.	.	.	.	.	.	.	.	.	50+	54	55	109	4.9
5.	.	.	.	.	.	-	-	.	.	42+	61	63 + 1*	125	5.6
6.	.	.	.	.	.	-	-	-	.	35+	79	81 + 1*	161	7.2
DELETIONS OF OTHER IDENTIFIERS														
7.	.	.	.	.	.	.	.	.	-	40+	45	46	91	4.1
8.	.	.	.	.	.	.	.	-	.	35+	52	51	103	4.6

* See Footnotes for Tables 2 and 3

[1] Based on using all 30,000 potentially linkable search records. Percentages are calculated on 2243 potential good links.

Key to Headings:
I = Initials of first (I1) or second (I2) given name
R = Remainder of first (R1) or second (R2) given name
MM = Mother's maiden surname
BP = Birth place

Key Entries:
Identifier: Deleted (-); Present (+); Present or absent (.)

Footnotes Regarding Tables 7 and 7a

1. * In Tables 7 and 7a the "false negatives" with an asterisk are due to potentially linkable search records becoming matched preferentially with the wrong death records; the weights for these will normally be below the "best" threshold so that they are unlikely to become "false negatives" that are due solely to the weight for a correct match falling below the "best" threshold.

2. The "best" threshold is taken to be the threshold at which the "false positives" and the "false negatives" are most nearly equal in number.

3. Of 30,000 miners' records, 2243 are judged to have been correctly matched with the corresponding death records as a result of the "hybrid" search. Of the 2243 correctly matched pairs, 705 had the "full" set of identifiers represented on both records in each of the pairs. Table 7 is based on the death records from the 705 pairs with full identifiers, and Table 7a is based on the 2,243 correctly matched pairs from the file of 30,000 potentially linkable search records.